AN INVESTIGATION OF THE POWER OF STOUT'S TEST OF ESSENTIAL UNIDIMENSIONALITY

By

CHENG ANG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1992

ACKNOWLEDGEMENTS

I wish to thank all my committee members for their valuable time, advice, and guidance during my doctoral study. My greatest appreciation and thanks go to Dr. James Algina and my committee chairman, Dr. David Miller. Their patience, kindness, encouragement, and insight helped make this dissertation possible. Also, many thanks go to Dr. Linda Crocker, who served as my temporary advisor during the first two semesters of my doctoral study, and to Dr. Clemens Hallman, who coordinated the Bilingual Education Fellowship that made my doctoral study possible.

I would like to express my appreciation to Dr. Anthony Lagreca and Dr. Stephen Golant for their encouragement and concern, and for giving me the opportunity to work with them as a graduate research assistant. Working with them has been one of the most enriching experiences of my academic career at the University of Florida. Last, but not least, my gratitude and love go to my wife, Rachel, for her love, support, and understanding, which made the pursuit of my doctoral degree possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

AN INVESTIGATION OF THE POWER OF STOUT'S TEST OF ESSENTIAL UNIDIMENSIONALITY

By

CHENG ANG

August 1992

Chairman: Dr. Michael Miller
Major Department: Foundations of Education

The power of Stout's procedure to detect deviations from essential unidimensionality in two-dimensional data was investigated for minor, moderate, and large deviations from unidimensionality. The criteria used in the categorization of deviations from unidimensionality were based on Shepard, Camilli, and Williams's categorization of area measures of item bias.

The power of Stout's procedure was directly related to the deviation from unidimensionality based on deviation areas. Deviation areas were inversely related to the correlation between the dominant ability and the reference composite. When the sample size increased, the power of Stout's procedure also increased. The power for 40-item tests was higher than for 20-item tests. When the proportion of items loaded on the minor dimension was 20%,

the power was the highest. Although the power for the 20% condition was higher than the 100% condition, the correlations $\rho_{\gamma,\theta 1}$ for the 20% condition were also extremely high. For the 10% and 20% conditions, even when the correlations $\rho_{\gamma,\theta 1}$ were near 1.00, the rejection rates were high.

In general, Stout's procedure had sufficient power to reject the null hypothesis of essential unidimensionality if 10% to 20% of the items were dimensionally distinct from the rest of the items. This is because only 10% to 20% of the items are being selected into the subtest (AT1) used in testing essential unidimensionality. When AT1 is dimensionally distinct from the rest of the items, Stout's null hypothesis of essential unidimensionality will be rejected.

The results of this study indicate that for minor deviation from unidimensionality, the rejection rates of Stout's procedure were not near the nominal level of 5%. For moderate and large deviation from unidimensionality, Stout's procedure had power to reject the null hypothesis of essential unidimensionality, especially if the sample size was 1,500 and the test length was 40. Further studies are recommended.

CHAPTER 1
INTRODUCTION

The purpose of the study was to investigate the power of Stout's procedure to detect deviations from essential unidimensionality in two-dimensional data. Included were data sets of minor, moderate, and large deviations from unidimensionality. Because item response theory was central to this study, an overview of this approach is appropriate. This chapter includes the following sections:

1. an overview of IRT models,
2. a statement of the problem,
3. the purpose of the study, and
4. a description of the significance of the study.

Overview of Item Response Theory (IRT)

Unidimensional IRT

One widely used class of test-theory models is the unidimensional IRT models (Lord, 1980). The attractiveness of the unidimensional IRT models in research and measurement is due primarily to the property of parameter invariance, which means that the item parameters are independent of the samples of examinees and that the person parameters are independent of the items included in the test. The well-

defined conditional standard errors expressed as functions of ability (θ) have also made IRT a very attractive class of models. Parameter invariance gives unidimensional IRT advantages over traditional test theory in solving many testing problems such as differential item functioning (DIF), equating, and tailored testing.

The commonly used unidimensional IRT models are the logistic and the normal ogive models. The three-parameter logistic (3PL) model can be written as follows:

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+\exp(-1.7a_i(\theta-b_i))}. \qquad (1.1)$$

In this model, b_i is the difficulty parameter, a_i is the discrimination parameter, and c_i is the lower asymptote. When the a parameter is set to a value and the c parameter is set to zero, the result is the one-parameter Rasch model.

$$P_i(\theta) = \frac{1}{1+\exp(-1.7a(\theta-b_i))}. \qquad (1.2)$$

Similarly, when only the c parameter is set to zero, the result is the two-parameter logistic (2PL) model.

$$P_i(\theta) = \frac{1}{1+\exp(-1.7a_i(\theta-b_i))}. \qquad (1.3)$$

The three-parameter normal ogive model is as follows:

$$P_i(\theta) = c_i + \frac{(1-c_i)}{(2\pi)^{1/2}} \int_{-\infty}^{a_i(\theta-b_i)} \exp(-\frac{y^2}{2})\, dy. \qquad (1.4)$$

As in the logistic model, a_i is the discrimination parameter, b_i is the difficulty parameter, and c_i is the

lower asymptote. One- and two-parameter normal ogive models are

$$P_i(\theta) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{a(\theta - b_i)} \exp\left(-\frac{y^2}{2}\right) dy \qquad (1.5)$$

and

$$P_i(\theta) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{a_i(\theta - b_i)} \exp\left(-\frac{y^2}{2}\right) dy \qquad (1.6)$$

respectively.

The logistic and normal ogive models are very similar. They have upper and lower asymptotes on their item characteristic curves. Both can have up to four parameters per item (the fourth is an upper asymptote which may be less than 1.00), but more commonly have three. The b_i parameter (item difficulty) is the point of inflection, the a_i parameter (discrimination) is the slope at the inflection point, and the c_i parameter (guessing parameter) is the lower asymptote. For computational convenience, logistic models are more widely used. The assumptions underlying unidimensional item response theory models are local independence conditional on a single ability and item characteristic curves (ICCs) that follow the logistic or normal ogive function.

Multidimensional IRT

Although several multidimensional IRT models have been proposed, most multidimensional IRT procedures are at the early stages of development and have not been widely used

because of computational and conceptual difficulties. Multidimensional IRT models with compensatory abilities and with noncompensatory abilities have been developed and described in the literature (Reckase, 1985). Compensatory models assume that high ability on one dimension compensates for low ability on another dimension in terms of the probability of a correct response; however, in noncompensatory models, high ability on one dimension does not compensate for low ability on another dimension (Ackerman, 1989). The compensatory model that is a multivariate extension of the 2PL model and has a common difficulty parameter (Reckase & McKinley, 1983) is

$$P_i(\theta_{jk}) = \frac{1}{1+\exp(d_i - \sum_{k=1}^{n} a_{ik}\theta_{jk})} \qquad (1.7)$$

where

$P_i(\theta_{jk})$ is the probability of a correct response by person j on item i,
θ_{jk} is the ability parameter for person j on dimension k,
a_{ik} is the discrimination parameter for item i on dimension k, and
d_i is the difficulty parameter for item i.

The multidimensional item difficulty (MID) is defined by

$$MID_i = -\frac{d_i}{(\sum_{k=1}^{n} a_{ik}^2)^{1/2}}, \qquad (1.8)$$

the multidimensional discrimination is defined by

$$MDISC_i = (\sum_{k=1}^{n} a_{ik}^2)^{1/2} = -\frac{d_i}{MID_i}, \quad (1.9)$$

and the correlation between item i and dimension k is defined by

$$\cos \alpha_{ik} = \frac{a_{ik}}{(\sum_{k=1}^{n} a_{ik}^2)^{1/2}}. \quad (1.10)$$

A compensatory model that is a multivariate extension of the 3PL model and has difficulty estimates for each dimension was defined by Reckase and McKinley (1983) as follows:

$$P_i(\theta_{jk}) = c_i + \frac{(1-c_i)}{1+\exp[-1.7\sum_{k=1}^{n} a_{ik}(\theta_{jk}-b_{ik})]} \quad (1.11)$$

where

$P_i(\theta_{jk})$ is the probability of a correct response by person j on item i,
θ_{jk} is the ability parameter for person j on dimension k,
a_{ik} is the discrimination parameter for item i for dimension k,
b_{ik} is the difficulty parameter for item i for dimension k, and
c_i is the lower asymptote for item i.

A multivariate extension of the noncompensatory 3PL model (Reckase & McKinley, 1983) is as follows:

$$P_i(\theta_{jk}) = c_i + \frac{(1-c_i)}{\prod_{k=1}^{n} [1+\exp(-1.7a_{ik}(\theta_{jk}-b_{ik}))]} \quad (1.12)$$

where all variables were defined as in equation 1.11. The compensatory model (equation 1.11) was used in the present

study to generate two-dimensional data as Reckase and McKinley (1983) considered it easier to work with and more practical.

Statement of the Problem

Unidimensionality means that the items in a test measure one ability or one dimension and that a single ability is sufficient to explain the examinee's performance. Although scores reported for achievement tests attempt to measure only a single ability, Harrison (1986) and Traub (1983) have argued that the assumption of unidimensionality may be problematic since every test and every response by examinees are multidimensional to some degree. Examinees are influenced by factors such as instructional emphasis, examinees' language comprehension, speed of completing the tests, and anxiety. The number of latent abilities that an individual uses to obtain a correct response may also vary from item to item (Traub, 1983). Because of multidimensionality, the use of unidimensional IRT models may be limited. To optimize the appropriate use of unidimensional IRT models, there is a need for a significance test to assess the dimensionality of test item data (Lord, 1980).

The multidimensional data found in most achievement tests consist of one major trait (dimension) of interest coupled with several minor traits. Typically, these minor

traits are unique to only a few items, or their influence on the items is relatively weak compared to that of the major dimension. In these cases, the minor traits should be ignored in the assessment of unidimensionality (Humphreys, 1981). Although over 80 indices to assess unidimensionality have been identified (Hattie, 1984), most assessment procedures do not make a distinction between the major and the minor dimensions. There has also been no widespread agreement among psychometricians as to the accuracy of these indices. Because minor traits should be ignored in the assessment of unidimensionality (Humphreys, 1986), it is important to have a hypothesis test procedure that is sensitive to a major dimension but excludes the minor dimensions.

Stout (1987) provided a significance test of essential unidimensionality in which the major dimension was counted but the minor dimensions were ignored. According to Stout (1987) and Nandakumar (1991), an essentially unidimensional data set should be appropriate for unidimensional IRT estimation; that is, when the weight of the minor dimension relative to the major dimension is small, the use of standard IRT data analysis procedures that require unidimensionality (for example BILOG or LOGIST) will work well. The power of Stout's procedure to reject essential unidimensionality may be important to practitioners who use Stout's procedure and who may need to decide if a set of

data is appropriate for unidimensional IRT estimation. It is therefore important to investigate the power of Stout's procedure in testing for essential unidimensionality.

Purpose of the Study

The purpose of this study was to investigate the power of Stout's procedure to detect deviations from essential unidimensionality in two-dimensional data. Included were data sets of minor, moderate, and large deviation from unidimensionality. The power of Stout's procedure was expected to be directly proportional to the deviations from unidimensionality. The specific questions were as follows:

1. How do minor, moderate, and large deviations from unidimensionality affect the power of Stout's procedure for testing essential unidimensionality?
2. How does sample size affect the power of Stout's procedure for testing essential unidimensionality?
3. How does test length affect the power of Stout's procedure for testing essential unidimensionality?
4. How does the proportion of items loaded on the minor dimension affect the power of Stout's procedure for testing essential unidimensionality?

Significance of the Study

An achievement test not only may measure the ability purported to be measured but also may be contaminated by one or more other abilities displayed by the person taking the test (Traub, 1983). The magnitude of this contamination may affect the estimation of the parameters of unidimensional IRT models and may create problems in the interpretation of test results. If the power of Stout's hypothesis test of essential unidimensionality is adequate, the test of essential unidimensionality may help determine the suitability of a set of multidimensional data for unidimensional IRT estimation. If Stout's procedure has adequate power to test essential unidimensionality, the use of unidimensional IRT models may be more accurate and the interpretation of test results may be more certain and precise. This will enhance the credibility of estimation of IRT models and applications of unidimensional IRT, such as equating, DIF studies, and adaptive testing.

CHAPTER 2
LITERATURE REVIEW

The purpose of this study was to investigate the power of Stout's procedure to detect deviations from essential unidimensionality in two-dimensional data; therefore, three topics are reviewed in this chapter:

1. the dimensionality assessment procedures,
2. the assumption of unidimensionality in IRT and the robustness of multidimensional data in unidimensional IRT models, and
3. essential unidimensionality and Stout's nonparametric procedure in assessing dimensionality.

Dimensionality Assessment Procedures

Although there is a great need for a test to assess unidimensionality (Lord, 1980) and over 80 indices to assess unidimensionality have been identified (Hattie, 1984), there has been no widespread agreement among psychometricians as to the accuracy of those indices. The indices identified by Hattie (1984) ranged from the unacceptable measures of answer pattern (e.g., scalogram analysis) to the more acceptable uses of latent traits and related statistics

(e.g., residual analysis). According to Hattie (1985), most procedures lacked a clear rationale and decision criteria for determining dimensionality, but the most widely used procedures have been the factor analytic procedures (Mislevy, 1986).

Factor Analysis

The traditional psychometric approach to the assessment of dimensionality has been factor analysis (Zwick, 1987). Both common factor analysis, with estimated communalities in the diagonal, and principal component analysis, with values of 1.00 in the diagonal, have been used and the outcomes have been similar (Reckase, 1981). Linear factor analysis of phi correlations is not suitable for assessing dimensionality because of a violation of the linearity assumption that leads to the introduction of difficulty factors (Mislevy, 1986). But, factor analysis of tetrachoric correlations, like IRT, assumes a latent continuous variable underlying each dichotomous variable, thus overcoming the problems of difficulty factors. The theoretical relationship between factor analysis of tetrachoric correlations and IRT has been well established (Bejar, 1983; Hulin, Drasgow, & Parsons, 1983, Chap. 8). The following common factor model is conceptually similar to IRT:

$$\gamma_i = \rho_i\theta+\varepsilon \qquad (2.1)$$

where

γ_i is the latent response to item i (underlying trait of each item) with mean zero and unit variance,
θ is the common latent trait measured by the items (common factor) and is normally distributed with mean zero and unit variance,
ρ_i is the correlation between γ_i and θ, and
ε is the error component equal to $(1-\rho_i^2)^{1/2}S$ where S is the specific factor which is uncorrelated with θ and has unit variance (Hulin et al., 1983).

The product moment correlation between γ_i and θ, ρ_i, is the factor loading of the jth item on the latent variable in factor analysis (Hulin et al., 1983). The parameters of the two-parameter normal ogive model of IRT,

$$P_i(\theta) = \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{a_i(\theta-b_i)} \exp(-\frac{y^2}{2})\,dy, \qquad (2.2)$$

can also be expressed in terms of the factor loading for both a_i (discrimination parameter) and b_i (difficulty parameter) as follows:

$$a_i = \frac{\rho_i}{(1-\rho_i^2)^{1/2}} \qquad (2.3)$$

and

$$b_i = \frac{\gamma^*}{\rho_i}. \qquad (2.4)$$

From equation (2.3), the factor loading for the common factor can also be derived (Hulin et al., 1983) as follows:

$$\rho_i = \frac{a_i}{(1-a_i^2)^{1/2}}. \qquad (2.5)$$

The γ^* is the threshold value at which examinees with γ_i ability would answer the item correctly if the ability

equaled or exceeded the threshold value γ^* (Bejar, 1983). Similarly, if the ability level is below the threshold, the item would not be answered correctly. As the threshold increases, the difficulty of the item increases.

Although factor analysis of tetrachoric correlations overcomes the problem of difficulty factors, it has the problem of creating artificial guessing factors. In addition, the correlation coefficient matrix may not be positive definite because the coefficients become unstable as their values approach +1 or -1 (Bock, Gibbons, & Muraki, 1988). Even with a smoothed positive definite matrix of tetrachoric correlations adjusted for guessing, guessing factors may still emerge (Lawrence & Dorans, 1987). Factor analysis of tetrachoric correlations is inappropriate for nonnormal distributions of ability, for item response functions that are not normal ogives, or when guessing is present. Also, tetrachoric correlations are subject to relatively large sampling errors (Lord, 1980).

An alternative procedure is to factor a phi correlation matrix with nonlinear factor analysis (McDonald, 1981). Factor analysis of responses to a set of binary items yields unidimensional results if and only if the data fit a nonlinear factor analysis model with one common factor. The nonlinear regression of binary variables on a common factor yields the item characteristic curves of IRT because "the regression of the binary items (score unity or zero) on a

common factor is the curve of the conditional probability of the item that is scored unity and is also known as the item characteristic curve" (McDonald, 1985, p. 231). Nonlinear factor analysis was one of the most accurate methods used by Hambleton and Rovinelli (1986) in assessing dimensionality, but an accepted criterion for determining the appropriate number of factors to retain a solution has not been identified. This procedure is limited to only one- and two-parameter normal ogive models and assumes ability is normally distributed (McDonald, 1967, 1981, 1982). An "easy to use" computer program is not readily available because of the complexities of nonlinear regression procedures.

Christofferson (1975) and Muthen (1978) developed a generalized least square (GLS) method that alleviates the problems encountered in factor analyzing tetrachoric and phi correlations. Muthen's solution was similar to Christofferson's but was computationally faster. GLS uses more of the information in the binary data compared to the tetrachoric solution and provides more consistent estimates. The major problem with the GLS procedure is that its use is limited to tests with only 25 items because of heavy computational complexities. Mislevy (1986) also noted that the cost of this method increases linearly with an increase in the number of factors extracted.

Although Reckase (1981), Mislevy (1986), Zwick (1987), and others have found factor analysis to be a suitable

procedure for assessing dimensionality, exploratory factor analysis has several limitations. These limitations include the lack of a statistical test for the number of factors, the inaccuracy of the criteria used for identifying the number of factors to rotate, and the likelihood that the unique variance might be negative (known as a Heywood case or improper solution). Lord and Novick (1968) argued that evidence of one common factor is not sufficient evidence for unidimensionality, and factor analysis often indicates the presence of more than one factor when the data are unidimensional (Hambleton & Rovinelli, 1986).

Full-Information Item Factor Analysis

To overcome the problems and limitations of the factor analysis of phi correlations, tetrachoric correlations, and GLS, Bock and Aitkin (1981) developed full-information item factor analysis (FIFA) that uses all the information available in the matrix of a dichotomously scored response pattern (Kingston & McKinley, 1988). Bock and Aitkin's FIFA is a dichotomous factor analysis that is based on multidimensional IRT (an extension of the two-parameter normal ogive model). When FIFA is used to extract one factor only, the resulting item parameter estimates are those of the two-parameter normal ogive model. Currently, TESTFACT is the commercially available computer program that runs FIFA.

FIFA can be developed by first assuming that underlying the response of examinee i on item j is

$$\gamma_{ij} = \sum_{k=1}^{n} \rho_{jk}\theta_{ki} + \varepsilon_{ij} \qquad (2.6)$$

where

- γ_{ij} is the latent response to item j by person i (underlying trait of each item) with mean of zero and variance of one,
- θ_{ki} represents the kth latent variable (common factor),
- ρ_{jk} is the factor loading of the jth item on the kth latent variable, and
- ε_{ij} is the error component with multivariate normal, mean zero, and covariance matrix I (Zwick, 1987).

If γ_{ij} exceeds the threshold value (γ^*) for the jth item, then examinees with γ_{ij} ability will answer the item correctly (Zwick, 1987). The observed score of examinee i on item j equals 1 if the response is correct and 0 if it is incorrect. The conditional probability for the ith examinee answering the jth item correctly with latent variable θ_i can be found in the multivariate generalization of the two-parameter normal ogive model of IRT (Zwick, 1987):

$$P(x_{ij}=1|\theta_i) = \frac{1}{(2\pi\sigma_j)^{1/2}} \int_{\gamma_j}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{y_{ij}-\sum_{k=1}^{n}\rho_{jk}\theta ki}{\sigma_j}\right)^2\right] dy_{ij} = F_j(\theta_i) \qquad (2.7a)$$

where

$$\sigma_{ij} = (1-\sum_{k=1}^{n} \rho_{jk}{}^2)^{1/2}. \qquad (2.7b)$$

FIFA is not limited by the number of items or length of the test (Muraki & Engelhard, 1985). This procedure allows the researcher to avoid the problems of extremely easy or difficult items and to correct for guessing effects using the multivariate generalization of the three-parameter normal ogive model. When FIFA adjusts for guessing, the model is

$$F^*_j(\theta_i) = c_j + (1-c_j) F_j(\theta_i). \qquad (2.8)$$

FIFA also corrects for missing items and smooths the tetrachoric correlation matrix to produce a positive definite correlation matrix. It can put constraints on item parameter estimates and it provides the likelihood ratio test of the statistical significance of additional factors. It allows for a stepwise factor analysis to be performed as follows: Using chi-squares, the first factor is compared with the second factor, then the first two factors are compared with the third factor, and so forth. Thus, chi-square goodness-of-fit indices are used to test if the added factors are statistically significant (Bock et al., 1988; Kingston & McKinley, 1988).

When applied to empirical data, FIFA has been found to work well (Lawrence & Dorans, 1987; Zwick, 1987). Nonnormality of the major factors, presence of minor factors, and the number of major factors have no adverse effects on FIFA (Salih, 1987). Compared to Muthen's GLS, FIFA can run more items, which is more practical for

achievement tests. The FIFA options for specifying different types of response functions, as well as the ability distribution, and the possibility of estimating the guessing parameters give FIFA an edge over GLS.

Because FIFA was recently developed, research and application have not been extensive. Although FIFA appears to be a viable procedure for assessing dimensionality, there are no clear decision criteria for determining the number of dimensions. The chi-square goodness-of-fit test for added factors is too powerful with large sample sizes (Bock et al., 1988).

Other Procedures

Residual analysis involves fitting a unidimensional IRT model to the test data, using the model parameter estimates to predict the item performance data, and then summarizing the discrepancies or residuals. Although Hambleton and Rovinelli (1986) found the results of residual analysis disappointing, Hattie (1984), as well as Berger and Knol (1990), found that fitting a two-parameter model and examining the residuals proved to be useful.

Bejar's procedure (1980) for assessing dimensionality has four steps:

1. Identify a subset of items that appear to measure a trait different from the trait measured by the the total test.

2. Analyze the items in the subtest with a three-parameter model.
3. Repeat the analysis (step 2) with the total set of items.
4. Compare the two sets of b-value estimates; they should be linearly related if the items are measuring the common traits.

Bejar's analysis is useful, but only when a test is divided into content categories (subtests) and is a reasonably long test (Hattie, 1985). Also, Bejar's procedure cannot ensure that the subtests examined are consistent with the structure underlying item responses. Hambleton and Rovinelli (1986) found Bejar's method inappropriate for assessing dimensionality.

Modified parallel analysis (MPA) is a data simulation method that employs both factor analysis and IRT in its computation. Although it has proven useful in some studies, it does not provide the best test for dimensionality, especially when guessing is involved (Hulin et al., 1983).

A procedure presented by Holland and Rosenbaum (1986), called Holland and Rosenbaum's Test of Unidimensionality, Monotonicity, and Conditional Independence, has also been used in assessing dimensionality. Zwick (1987) and Ben-Simon and Cohen (1990), among others, have demonstrated the application of Holland and Rosenbaum's approach to assessing dimensionality. Although Zwick (1987) found that the

procedure yields results consistent with those of other procedures such as FIFA, Ben-Simon and Cohen (1990) found results obtained from Holland and Rosenbaum's approach were inaccurate.

Unidimensionality

According to Hambleton and Swaminathan (1985), when there are k latent traits or abilities that underlie examinee performance on a set of test items, these k latent traits or abilities define the k dimensions of a test. When a single ability or trait is assumed to account for examinee test performance, the test is called unidimensional. According to McDonald (1982), a meaningful definition of unidimensionality should be based on the principle of local independence, where for examinees with the same ability, the covariation between items in the test is 0.00. Hambleton and Swaminathan (1985) went further to state that the assumption of unidimensionality is equivalent to the assumption of local independence. Local independence requires that scores on any two items be uncorrelated (statistically independent) when the ability is fixed; that is, only one ability is necessary to account for the relationship among a set of test items. The principle of local independence can be expressed as

$$P(\underline{U}|\theta) = \prod_{i=1}^{n} [P_i(\theta)]^{u_i} [1-P_i(\theta)]^{1-u_i} \qquad (2.9)$$

where

$P(\underline{U}|\theta)$ is the probability of the response pattern conditional upon the latent response variable,

$P_i(\theta)$ is the probability of responding correctly to item i conditional upon θ,

$\underline{U}$ is the random item response pattern vector,

u_i is the binary random response variable on the ith item, and

θ is the k dimensional random vector of latent traits.

In other words, "The probability of the response pattern for each examinee is equal to the product of the probability associated with the examinee response to each item" (Hambleton & Swaminathan, 1985, p. 23).

Unidimensionality is neither time nor population invariant; a test may be unidimensional for a particular time or population but not for another time or population (Hambleton & Swaminathan, 1985). Thus, the construction of a unidimensional test must be time-based and, as much as possible, population-based. Although unidimensionality is very important, there is no widespread agreement on the effectiveness of procedures for assessing it (Lord, 1980).

The assumption of unidimensionality in most IRT models can be problematic because almost every test and every response by examinees are multidimensional; that is, examinees are influenced by instructional emphasis, language comprehension, speed of completing the tests, and anxiety. The number of latent abilities that an individual uses to

obtain a correct response may also vary from item to item (Traub, 1983). Folk and Green (1989) showed that the use of a unidimensional model with multidimensional data can bias parameter estimation, adaptive item selection, and ability estimation for both adaptive and nonadaptive tests. The uncritical use of unidimensional models, where multidimensionality holds, can have serious consequences (Reckase, Ackerman, & Carlson, 1988), especially if the abilities measured by the test were weakly correlated in the examinee group. McKinley (1983) found that a multidimensional model more accurately modeled the data than the unidimensional model with some data. Thus a unidimensional model should not be used if doing so leads to adverse effects on parameter estimation, item selection, or ability estimation.

Although item response theory assumes unidimensionality, Ackerman (1989) found that as the correlation between the generated two-dimensional abilities increased, the response data appeared to become more unidimensional. Harrison (1986) and Drasgow and Parsons (1983) have shown IRT to be robust to the unidimensionality assumption when the correlations among the dimensions are higher than .65. When a test is formed by items that all measure the same composite of abilities, that test will also meet the IRT requirements of unidimensionality; that is, the unidimensionality assumption only requires that items in a

test measure the same composite of abilties rather than a single ability (Reckase, Ackerman, & Carlson, 1988). Robustness to the unidimensionality assumption means that IRT is less restrictive than traditionally assumed.

Essential Unidimensionality

Although most reported scores from achievement tests are designed to measure a single trait to optimize the interpretation of the test, in practice most items are inherently multidimensional where item responses are a function of one major trait (dimension) of interest combined with several minor traits. These minor traits are unique only to a few items; they are not detrimental to the measurement of the dominant dimension or the assessment of dimensionality (Humphreys, 1986). Lord and Novick (1968) also referred to the complete latent space as consisting of one major and several minor factors. Unfortunately, the IRT definition based on the principle of local independence makes no distinction between major and minor dimensions.

Although only dominant traits should determine dimensionality, there was no precise definition of a dominant dimension until Stout (1987). He termed the major dominant dimension, with minor dimensions that can be ignored, essential unidimensionality. Stout (1987) replaced the usual assumption of unidimensionality with a weaker and a statistically testable assumption of essential

unidimensionality. Stout's definition of essential unidimensionality (1990, p. 299) is as follows:

> The essential dimensionality (d_E) of a test $\{U_i, i\geq 1\}$ is the minimal dimensionality required for a latent trait θ to make the latent model $\{U_N, \theta, N\geq 1\}$ an essential independent (EI), weakly monotone (WM) model. When $d_E = 1$, essential unidimensionality is said to hold. If essential dimensionality holds using ability θ then $\{U_i, i\geq 1\}$ is said to be essentially d_E dimensional with respect to ability θ. Such a trait is called an essential trait for $\{U_i, i\geq 1\}$

where

U_i is the infinite item pool,
U_N is the first N items of U_i making up the test U_N,
θ is the latent random vector, and
$\boldsymbol{\theta}$ is a particular value of θ.

Stout (1990, p. 297) defined essential independence by using the nonsparse subtest of the infinite item test framework as follows:

> The latent model $\{U_N, \theta, N\geq 1\}$ is said to be essentially independent (EI) if for every collection of nonsparse subtests for each θ in the range of θ,

$$D_N(\theta) = \binom{M(N)}{2}^{-1} \sum_{i,j\in I_N, i<j} Cov(U_i, U_j | \Theta=\theta) \to 0 \text{ as } N \to \infty \qquad (2.10)$$

where

I_n is the item pools of the subtest, and
$M(N)$ is the length of the subtest.

(Note: a subtest is nonsparse if it is nested ($\mu_N \subset \mu_{N+1}$) and there exist $\epsilon > 0$ such that $M(N)/N \geq \epsilon$ for all $N \geq 1$, where μ_N denotes a subtest. See Stout (1990, p. 297)).

Equation 2.10 also implies that for any collection of nonsparse subtests for each θ

$$D_N(\theta) = \binom{N}{2}^{-1} \sum_{1 \le i < j \le N} Cov(U_i, U_j | \Theta = \theta) \to 0 \text{ as } N \to \infty. \quad (2.11)$$

Using the infinite item test framework, Stout (1990) argued that the usual assumption of local independence, where the covariation in the set of items is 0.00 for examinees with the same ability, be replaced by an assumption of essential independence (EI) in which the covariation approaches 0.00 as N for items approaching infinity; that is, on the average, the conditional covariances should be small, and local independence should hold approximately.

Instead of the monotonic increasing property of an ICC, Stout (1990) suggested the ICC could be weakly monotonic. The latent model (U_N, θ, $N \geq 1$) is said to be weakly monotone if

$$\sum_{i=1}^{N} P_i(\theta)$$

is nondecreasing in θ, for all $N \geq N_o$ for some fixed N_o. A test conforming to the following conditions is essentially unidimensional:

(1) Few items of U_N depend on an ability (or abilities) other than the ability of interest,
(2) each ability other than the ability of interest influences at most a small number of items of U_N and moreover these incidental abilities are "orthogonal" to each other, conditional on the ability of interest, and
(3) the magnitude of the dependence of the items of U_N on an ability (or abilities) other than the ability of interest is small, even though most of the items may depend on this other ability. (Stout, 1990, p. 300)

Based on infinite item pools, Stout (1990) has also shown mathematically how essential unidimensionality may or may not fail (e.g., 2^k-th of infinite items are essential unidimensional, but 10k-th are not).

Foundation of Stout's Procedure

Stout's (1987) nonparametric procedure for assessing dimensionality is based on his definition of essential dimensionality and on essential independence in particular. Let

$$\overline{U} = \frac{1}{N} \sum_{i=1}^{N} U_i \tag{2.12}$$

denote the proportion correct, and let

$$Var_d(\overline{U}|\Theta_1=\theta_1) = E[\overline{U}^2|\Theta_1=\theta_1] - [\overline{U}|\Theta_1=\theta_1]^2 \tag{2.13}$$

denote the variance in $\overline{U}$ due to $\Theta_2,\ldots\Theta_d$ at fixed $\Theta_1 = \theta_1$ for the minor dimensions.

Let

$$Var_1(\overline{U}|\Theta_1=\theta_1) = \frac{1}{N^2} \sum_{i=1}^{N} Var(U_i|\Theta_1=\theta_1] \tag{2.14}$$

denote the variance in $\overline{U}$ at a fixed $\Theta_1 = \theta_1$ (major dimension) under the assumption that $U_1,\ldots,U_N$ are locally independent with respect to the one dimensional trait Θ_1. Stout (1990) showed that

$$Var_d(\bar{U}|\Theta_1=\theta_1) = \frac{N-1}{N} D_N(\theta_1) + Var(\bar{U}_i|\Theta_1=\theta_1]. \qquad (2.15)$$

At fixed θ_1, ((N-1)/N) $D_N(\theta_1)$ (the covariation between items) is equal to the difference between the true variation in student scores at $\theta_1 = \theta_1$, taking into account the minor dimensions ($\theta_2,....\theta_d$), and the estimated one dimensional variance at $\theta_1 = \theta_1$, ignoring the nuisance dimensions. If $D_N(\theta_1)$ approaches zero, then the nuisance dimensions $\theta_2,....\theta_d$ do not contribute significantly to the variance in $\bar{U}$ and the item pool is essentially unidimensional (Nandakumar, 1987).

For finite length tests, a test of length $U_1,U_2...U_N$ is essentially unidimensional if

$$S_{M,N} = \frac{1}{M(M-1)} \sum_{1\le i\ne j\le N} Cov(U_i, U_j|Y_p) \approx 0 \qquad (2.16)$$

where M is the subtest $U_1,U_2...U_M$ of length M (M<N) and Y_p is the proportion correct on the long subtest complementary to $U_1,U_2...U_M$ with length n = N - M. This equation suggests splitting the test into two subtests such that if $S_{M,N}$ is approximately equaled to zero, the test is essential unidimensional (Nandakumar, 1987; Stout, 1987).

Stout's Procedure

Stout (1987) first developed a procedure for assessing essential unidimensionality. It has been further refined by Nandakumar and Stout (in press) to adjust for guessing in the presence of high discriminations. Stout's procedure is

nonparametric and is consistent with the concept of dimensionality used in factor analysis, which is sensitive to the dominant trait or traits but not sensitive to minor traits. The procedure described here is based on a sample size of less than 2000. For a sample size larger than 2000, minor modifications in the test statistics are necessary, and these have been described in Stout (1987). Stout, Nandakumar, Junker, Chang, and Steidinger (1991) have prepared a computer program (named DIMTEST) to test for essential dimensionality of a set of data. The program includes all refinements by Nandakumar (1987) and Nandakumar and Stout (in press). A summary of Stout's procedure based on Stout (1987, 1990), Nandakumar (1987, 1991), and Nandakumar and Stout (in press) is as follows:

Step 1. Divide the N test items to be analyzed into three subtests, two short assessment subtests (AT1 and AT2) of length M each, and a long partitioning subtest (PT) of length n. The subtest AT1 is selected with M items ($M < N/4$) that are as unidimensional as possible and as dimensionally distinct from the rest of the test as possible. This can be achieved in one of two ways: (a) using experts' opinions to select M items of the same content area, or (b) using factor analysis of tetrachoric correlations to select M items with the highest loadings of the same sign on the second extracted factor. The scores on AT1 are used to compute Stout's statistics. Next, M items

of AT2 are selected so that the difficulty levels of these items are as similar to those of the AT1 items as possible. In addition, the content of the AT2 items are a representative sample of items from PT, where PT are the remaining items n= N-2M items. The subtest AT2 is thus dimensionally similar to PT and, at the same time, its difficulty distribution is similar to AT1's. The scores on AT2 are used to correct for the pre-asymptotic statistical bias in Stout's statistics.

Step 2. Assign examinees with the same PT scores to the same subgroup so that each subgroup consists of examinees with approximately equal ability. The purpose of PT is to group examinees into subgroups. When the test is essentially unidimensional within each subgroup, examinees are assumed to be approximately of equal ability. Examinees getting all the PT items right or wrong are excluded. If n=16 there will be 15 subgroups. In general, each subgroup must be composed of at least 20 examinees to maintain agreement with asymptotic theory. Subgroups with too few examinees are deleted.

Step 3. Compute the "usual" variance estimates for the kth subgroup of examinee on AT1.

$$\hat{\sigma}_k^2 = \sum_{j=1}^{J_k} \frac{(Y_j^{(k)} - \overline{Y}^k)^2}{J_k} \qquad (2.17)$$

where

$$Y_j^{(k)} = \sum_{i=1}^{M} \frac{U_{ijk}}{M}, \tag{2.18}$$

$$\overline{Y}_j^{(k)} = \sum_{j=1}^{J_k} \frac{Y_j^{(k)}}{J_k}, \tag{2.19}$$

and where

U_{ijk} is the response of the jth examinee to the ith item from subgroup k,
$J_{k(k)}$ is the number of examinees in subgroup k of PT,
$Y_j^{(k)}$ is the AT1 score of the jth examinee from subgroup k, and
$\overline{Y}_j^{(k)}$ is the average examinee AT1 score for subgroup k.

Step 4. Compute the unidimensional variance estimate for the kth subgroup in AT1.

$$\hat{\sigma}^2_{U,k} = \sum_{i=1}^{M} \frac{\hat{P}_i^{(k)}(1 - \hat{P}_i^{(k)})}{M^2} \tag{2.20}$$

where

$$\hat{P}_i^{(k)} = \sum_{j=1}^{J_k} \frac{U_{ijk}}{J_k}. \tag{2.21}$$

Step 5. Combine and normalize the different subgroup variance estimates to form the statistics T_L for AT1.

$$T_L = \frac{1}{k^{1/2}} \sum_{k=1}^{k} \left[\frac{\hat{\sigma}_k^2 - \hat{\sigma}^2_{U,k}}{S_k} \right] \tag{2.22}$$

where L stands for long test and

$$S_k^2 = \frac{[(\hat{\mu}_{4,k} - \hat{\sigma}_k^4) + \frac{\hat{\sigma}_{4,k}}{M^4}] + 2\sqrt{(\hat{\mu}_{4,k} - \hat{\sigma}_k^4)\frac{\hat{\sigma}_{4,k}}{M^4}}}{J_k}, \tag{2.23}$$

$$\hat{\mu}_{4,k} = \sum_{j=1}^{J_k} \frac{(Y_j^{(k)} - \overline{Y}^k)^4}{J_k}, \tag{2.24}$$

and

$$\hat{\sigma}_{4,k} = \sum_{i=1}^{M} \hat{P}_i^{(k)} (1 - \hat{P}_i^{(k)}) (1 - 2\hat{P}_i^{(k)})^2. \tag{2.25}$$

Step 6. Similarly, compute the statistic T_B on AT2 by repeating steps 3 to 5 on items in AT2 and compute T_B according to the equation given for T_L.

Step 7. Perform the test for unidimensionality by Stout's statistic T given by

$$T = \frac{(T_L - T_B)}{\sqrt{2}}. \tag{2.26}$$

The hypothesis to be tested is as follows:

H_o: d = 1 versus H_1: d > 1

where d denotes the essential dimensionality of the latent space. H_o is rejected if $T \geq Z_\alpha$ where Z_α is the upper 100(1-α) percentile for the standard normal distribution, and α is the desired level of significance (Nandakumar, 1987, 1991; Stout, 1987, 1990).

Stout's test statistic is based on two variance estimates (Nandakumar, 1987; Stout, 1990): the usual variance estimate ($\hat{\sigma}^2_k$), which is sensitive to multidimensionality, and the unidimensional variance estimate ($\hat{\sigma}^2_{U,k}$), which is not sensitive to multidimensionality. If local independence holds, that is,

if the data are unidimensional, the two variance estimates will be equal. If the two estimates vary widely, the data may be judged to be multidimensional. The comparison of variance estimates is basically accomplished through the AT1, AT2, and PT partition. If a test is unidimensional, use of the PT scores to ensure that examinees within each subgroup are approximately of equal ability leads to approximately equal variance estimates based on AT1. If the test is multidimensional, the PT would have items testing different abilities, but items in AT1 would still test the same ability. In this case, usual variance estimates will be relatively large, but unidimensional variance estimates would still be smaller, resulting in a significant test statistic.

Application of Stout's Procedure

According to Nandakumar (1991, p.104), a test that is "approximately essential unidimensional means that the use of standard IRT data analysis procedures (for example, BILOG or LOGIST) that require unidimensionality will work well." Therefore, Stout's hypothesis test of essential unidimensionality may be used for assessing the appropriateness of a set of data for unidimensional IRT estimation. In addition to testing for essential unidimensionality, Stout's procedure may be used for the construction of an essentially unidimensional test or to decide whether two-dimensional tests administered to the

same population are measuring the same thing (Stout, 1987).

There are many advantages to Stout's procedure. It is appropriate for assessing the dimensionality of test data, even if guessing is present (Nandakumar & Stout, in press). The program DIMTEST adjusts for guessing in the presence of high discriminations and provides for automatic selection of AT1 items. Its computational requirement (CPU) is modest compared to full information item factor analysis, Muthen's GLS, and other factor analytic procedures. The procedure is designed to be insensitive to the presence of minor dimensions because the influence is not viewed as important. Use of this procedure is supported by both asymptotic theory and a Monte Carlo simulation study (Nandakumar, 1991; Stout, 1987). Also, because the procedure is based on a nonparametric model, issues of parametric model correctness are avoided (Stout, 1987).

Simulation Studies Based on Stout's Procedure

The power of Stout's procedure had been studied by Stout (1987) and Nandakumar (1991). In Stout's study (1987), the item-parameters and test length simulated were based on samples of actual tests and were not systematically controlled. The numbers of examinees studied were 750, 2,000 and 20,000. When the data were strictly unidimensional (d=1), the rejection rates were $\leq$ 3% at the 0.01 significance level, $\leq$ 6% at the 0.05 significance level and $\leq$ 17% at the 0.10 significance level. The rejection

rates were slightly higher for the 2,000 examinee sample size compared to rejection rates for the 750 examinee sample size, but when examinee sample size was 20,000, the rejection rate was zero at the 0.05 significance level. For two-dimensional data with equally weighted dimensions, Stout (1987) found that the rejection rates decreased as the correlation (ρ) between the dimensions increased and the rejection rates increased when the sample size increased. For example, with a sample size of 750 and $\rho = 0.5$, the rejection rates were about 60%, but when $\rho = 0.7$, the rejection rates decreased to about 36%. Similarly, when the sample size was 2,000 and $\rho = 0.5$, the rejection rates were about 95%, but when $\rho = 0.7$, the rejection rates were all below 95%. For a sample size of 20,000, there was a 100% rejection rate at $\rho = 0.6$ and a 94% rejection rate at $\rho = 0.8$, showing good power at this large sample size. Stout (1987) also found that the presence of guessing ($c = 0.2$) lowered the rejection rates compared to the rejection rates determined in the absence of guessing ($c = 0.00$). For two dimensional data, on the average, Stout's procedure had moderate statistical power with an average of 81% rejection rates.

Nandakumar (1991) used the same item parameters, test length, and number of examinees (except 20,000) as used by Stout (1987) but varied the weight (ξ) of the second dimension (minor dimension) relative to the first dimension

(major dimension). Two values of ξ were studied: 0.2 and 0.4. The number of minor abilities also varied; they were determined by the number of items each minor ability influenced. Two cases were studied. For Case 1, one major ability with two combinations of minor abilities were studied; for scenario one, each minor ability influenced two items and for scenario two, each minor ability influenced five items. When each minor ability influenced two items, the total number of minor abilities influencing the items equaled the total number of items divided by 2. For Case 2, one major and one minor ability influenced all items in the test. Nandakumar (1991) also provided an index of deviation from essential unidimensionality, β = 15 * min [var (a_1), (a_2)]. Nandakumar (1991) showed that when each minor ability contributed to very few items (e.g., 2), the test was assessed as unidimensional, even when β was large, but when each minor ability contributed to many items (e.g., 5), Type I error was inflated. Although Nandakumar (1991) also showed an increase in rejection rates directly proportional to β for Case 1, test length was not being taken into account (as β decreased, test length increased), and the number of minor abilities varied with β (Nandakumar, 1991). For Case 2, Nandakumar (1991) showed that the rejection rates of Stout's procedure increased from ξ= .2 to .4. Nandakumar (1991) also noticed that the rejection rates were slightly higher for 50-item tests compared to 25-item tests.

However, the comparison of β was made without holding the item parameters constant; that is, different item parameters were used for different sizes of β. Although Nandakumar (1991) used β to provide a rough index in determining the power of Stout's procedure, the use of β as an index may be questionable because it is solely a function of the smaller of the variances a_1 and a_2. Unless β was extremely large (Nandakumar, 1991), there was no direct relationship between β and the rejection rate.

Stout's study (1987) of two-dimensional data included two major abilities; some items depended only on one ability and others depended on two abilities. Nandakumar's study of two-dimensional data (1991) was based on either one major and one minor ability or one major and many minor abilities. No research was done for the case in which a small proportion of items loaded on the second dimension, and only one minor ability influenced a small proportion of items on the second dimension.

Although Nandakumar (1991) noticed a slight increase in the rejection rates for longer tests with large β, test length has not been systematically studied. In Stout's preliminary investigation (1987), he found unacceptably large positive bias (Type I error) when the test length was short. Although he used the AT2 subtest to correct for the positive bias caused by short tests, no systematic study was

done on the impact of test length on the power of Stout's procedure.

Although the power of Stout's procedure has been studied (Nandakumar, 1991; Stout, 1987), all conditions manipulated were not conducted with known minor, moderate, and large deviations from unidimensionality. An index of deviation from unidimensionality may also be needed to guide the decision about the appropriateness of data for unidimensional IRT estimation.

Summary

This chapter has presented three major issues related to assessing essential unidimensionality. First, an overview of commonly used procedures in assessing dimensionality was presented with an emphasis on factor analytic procedures. Factor analysis, although it has been widely used, has many associated problems. FIFA has been shown empirically to be a good procedure in assessing dimensionality (Zwick, 1987), but using a chi-square test of the added factor for determining the number of dimensions may be too powerful with large sample sizes. Other commonly used procedures such as residual analysis, Bejar's procedure, MPA, and Holland and Rosenbaum's Test of Unidimensionality were also described, but evidence of their appropriateness in assessing dimensionality has not been conclusive.

Second, the unidimensionality assumption in IRT and the robustness of IRT models to violations of the unidimensionality assumption were reviewed. The unidimensional IRT models may be robust to multidimensionality if the items measured are the same composite of abilities (direction) or if the dimensions are highly correlated.

Finally, Stout's definition of essential unidimensionality and his non-parametric procedure for testing essential unidimensionality were reviewed. Although the power of Stout's procedure has been studied, it was not examined with known minor, moderate, and large deviations from unidimensionality. Test length and a small proportion of items on the second dimension have also not been studied as factors that may influence the power of Stout's procedure.

CHAPTER 3
METHODOLOGY

Four topics are addressed in this chapter:

1. purpose and research questions,
2. design of the study,
3. simulation models, and
4. analysis of data.

Purpose and Research Questions

The purpose of this study was to investigate the power of Stout's procedure to detect deviations from essential unidimensionality in two-dimensional data. Included were data sets of minor, moderate, and large deviations from unidimensionality. The power of Stout's procedure should be directly proportional to the deviations from unidimensionality. The specific questions were as follows:

1. How do minor, moderate, and large deviations from unidimensionality affect the power of Stout's procedure for testing essential unidimensionality?
2. How does sample size affect the power of Stout's procedure for testing essential unidimensionality?

3. How does test length affect the power of Stout's procedure for testing essential unidimensionality?
4. How does the proportion of items loaded on the minor dimension affect the power of Stout's procedure for testing essential unidimensionality?

Design of the Study

Test Length and Sample Size

In the present study, the test length and sample size each have two levels. The test lengths studied were 20 (a short test) and 40 (an average-length test). Small and large sample sizes of 700 and 1,500 were studied. Hattie (1984) stated that sample sizes smaller than 300 tended to be unstable for latent trait procedures. In addition, large sample sizes (>5,000) might result in inappropriate rejection rates (Stout, personal communication).

Item-Parameters

The item test parameters of the two-dimensional model in equation 1.11 with one major and one minor dimension were used in generating item response data. The first dimension was the major dimension that the test was purported to measure and the second dimension was the minor dimension. The influence of the second dimension on each item was relatively weak compared to that of the first dimension.

The means and variances of a_1 and a_2 reflected the degree to which their respective traits influenced item scores. An item with a large a_1 and a small a_2 was much more heavily influenced by θ_1 than by θ_2 and vice-versa (Nandakumar, 1991). Both θ_1 and θ_2 were normal with mean zero and variance equal to one. The correlation between the abilities was set at zero.

A preliminary investigation using Nandakumar's ξ (1991) to control the weight of the major dimension relative to the weight of the minor dimension was carried out. The item test parameters a_1 and a_2 were computed by varying μ, σ and ξ in the following expressions:

$$a_1 \sim N((1-\xi)\mu,\ (1-\xi)^{1/2}\sigma)$$

$$a_2 \sim N(\xi\mu,\ \xi^{1/2}\sigma)$$

$$a_1 + a_2 \sim N(\mu,\ \sigma) \qquad (3.1)$$

With $\mu = 1.07$ and $\sigma^2 = 0.16$, Stout's test of essential unidimensionality had little power. Even when the weight of the minor dimension was the same as the major dimension ($\xi = 0.5$), the rejection rates were less than 20%. With $\sigma^2 = 0.64$ however, there was substantial power even at $\xi = 0.3$. Using Nandakumar's ξ requires a large σ^2 because increasing ξ with a constant μ and σ^2 led to a decrease in $\sigma_{a1} = (1-\xi)^{1/2}\sigma$ and $\mu_{a1} = (1-\xi)\mu$. Unless σ_{a1}^2 is relatively large, the mininum of the variance of a_1 and a_2 will also be small (β) and lead to little power (Nandakumar, 1991). To avoid using a large σ^2 and hence a large range, in this study a small σ^2

was studied and the effect of the reduction in σ_{a1} and μ_{a1} (due to ξ) was controlled by holding both σ_{a1} and μ_{a1} constant.

In this study, the values of a_1 were fixed across conditions; that is, only one set of a_1 was used across deviation areas for each test length. For the 40-item tests, the a_1 parameters used in this study were the discrimination parameters of a 40-item ACT math test reported by Drasgow (1987). The mean and sigma of a_1 were 1.09 and 0.35, and a_1 ranged from .40 to 2.00. For the 20-item test, a_1 parameters were selected from the 40-item test parameters with mean 1.09, sigma 0.36 and a_1 ranged from 0.40 to 2.00. The mean and sigma for a_2 were W(1.09) and $W^{1/2}(0.35)$, where W is the weighting factor similar to Nandakumar's (1991) use of ξ (ξ = (W / 1 + W)); that is, the μ and σ of a_2 were weighted by $W\mu$ and $W^{1/2}\sigma$ of a_1 instead of $\xi\mu$ and $\xi^{1/2}\sigma$ of the common a_s in Nandakumar (1991). Although W and ξ are basically the same, the purpose of using W was to keep a_1 the same across deviation areas. Because a_1 did not change, a_1 and a_2 were equal when the weight W = 1.00 (as opposed to Nandakumar's ξ = 0.5).

To compute the parameters for a_2, the parameters for a_1 were randomly rearranged using random numbers for both the 40- and 20-item tests. The purpose of rearranging the values of a_1 was to use the new a_1 for computing the weighted a_2 so that a_2 would be statistically independent of

the original a_1 (the original a_1 was used for the major dimension). The item parameters of a_2 were computed by varying W on the new a_1 (e.g., a_2 = 0.34 * new a_1). W ranged from .34 to 0.90 to explore the desired deviation areas (this will be described further under the deviation areas section). Because only one set of random numbers was used for each test length to generate the new a_1, the item parameter a_2 for each item will have the same value across deviation areas if multiplied by 1/W.

The difficulty parameters reported by Drasgow (1987) for the ACT math test were also used in this study. The values reported by Drasgow were used for both b_1 and b_2 (b_1 has the same value as b_2). Because Drasgow (1987) only reported item difficulties for a 40-item test, b_1 and b_2 for the 20-item tests were selected from parameters for the 40-item test. The mean and standard deviation for b_1 and b_2 were about the same for the 20-item and the 40-item test parameters: μ of b_1 and b_2 were 0.50 and σ of b_1 and b_2 were 0.61 for both test lengths. The range, however, differed: for the 40-item test parameters, the range was from -1.02 to 1.50; for the 20-item test, the range was from -.60 to 1.50. For each test length, the same values of b_1 and b_2 were used across deviation areas to ensure that variation in deviation areas was not confounded with fluctuations in the difficulty parameters.

Proportion of Items

The proportion of items loaded on the minor dimensions had three levels: 10%, 20% and 100%. The 10% and 20% levels were studied because many achievement tests have 5% to 29% of the items loaded on the second dimension (Ackerman, 1987). For example, a math test might consist of 10% word problems, thus requiring the ability to comprehend sentence structure. The 100% level was studied because it is common to have all test items contaminated by a second trait, although the contamination might be relatively weak compared to the influence of the first dimension (Nandakumar, 1991). For example, the ability of an examinee to answer all the math test items might be influenced by the examinee's ability to understand the instructions in English.

For the three proportions of items loaded on the second dimension, the parameters (a_1 and b_1) for the major dimension were the same across deviation areas for each test length. For the minor dimension, when 10% and 20% of the items loaded on the minor dimension, those items had the same a_2 and b_2 as some matched items in the 100% condition; that is, 10% and 20% of the items were selected from the 100% condition and the rest of the item loadings on the minor dimensions were set to 0.00. The selection of items for the 10% and 20% of the items loaded on the minor dimension will be discussed further under the deviation areas section.

Analytical Estimates

Equations for estimating the unidimensional item test parameters of the two-parameter model from the trait and item test parameters of a two-dimensional compensatory model have been established by Wang (1986). Because the data in this study were generated based on a bivariate extension of the 2PL model with compensatory abilities (equation 3.5) and the dimensions were assumed to be uncorrelated, Wang's (1986) special case formula was used in the estimation of the parameters of the unidimensional two-parameter model:

$$\hat{a}_j = \frac{w_1' a_j}{\sqrt{(1 + a_j' w_2 \; w_2' a_j)}} \tag{3.2}$$

and

$$\hat{b}_j = \frac{b_j \sqrt{a_j' a_j}}{w_1' a_j} \tag{3.3}$$

where w_1(p x 1) is the first eigenvector of the matrix A'A;
w_2(p x 1) is the second eigenvector of the matrix A'A;
A(n x p) is the matrix of p discrimination parameters for each of n items;
a_j(1 x p) is the ith row of A, a vector of discriminations for item j; and
b_j(1 x p) is the ith row of b, a vector of difficulies for item j.

Categorization of Deviations

After the parameters of the unidimensional model were analytically estimated using Wang's procedure, differences between the analytical estimations and the first dimension of the true parameter values were computed using the

unsigned area (UA) between the two item characteristic curves (ICCs) (Raju, 1988). The UA was computed by using

$$UA = \left| \frac{2(\hat{a} - a_1)}{Da_1\hat{a}} \ln\left(1 + \exp\left(\frac{Da_1\hat{a}(\hat{b} - b_1)}{(\hat{a} - a_1)}\right)\right) - (\hat{b} - b_1) \right| \quad (3.4)$$

where

a_1 is the discrimination parameter for the major dimension,

$\hat{a}$ is the discrimination parameter for the estimated dimension,

b_1 is the difficulty parameter for the major dimension, and

$\hat{b}$ is the difficulty parameter for the estimated dimension.

The area was then averaged over all the items loaded with two dimensions for each test. The deviations were grouped into three categories based on the average area: minor, moderate, and large deviations.

The criteria used to determine the three levels of deviations were based on the criteria used by Shepard, Camilli, and Williams (1985) in the categorization of bias between two groups. Their criteria were based on the differences between the difficulty parameters, b_1-b_2, of the two groups. When b1-b2 was less than .20, an item was classified as unbiased; when b1-b2 was between .20 and .35, an item was classified as weakly biased; and when b1-b2 was greater than .35, an item was classified as moderately biased. Their rationale for the categorization of biases was based on the examination of actual data (Shepard et al., 1985).

Since Raju's (1988) area procedure for the Rasch model between two ICCs was UA = $|b_1-b_2|$, the absolute value of the index used by Shepard et al. (1985) would be equivalent to that of Raju's area. Thus, in this study, when the area was less than .20, it was classified as a minor deviation; an area between .20 and .35 was classified as a moderate deviation; and an area greater than .35 was classified as a large deviation. Table 3.1 shows the characterizations of the three categories of deviations from unidimensionality.

Table 3.1

Three Deviations from Unidimensionality and Criteria for Cut-Off

Raju's UA	Deviation Category
< .20	Minor
≥ .20 ≤ .35	Moderate
> .35	Large

From the three categories of deviations from unidimensionality, six unique areas were chosen for the generation of data and testing of the hypothesis of essential unidimensionality. The areas chosen were 0.19 for a minor deviation; 0.28, 0.31 and 0.34 for a moderate deviation; 0.37 and 0.40 for a large deviation. The area of 0.19 represented the maximum area for a minor deviation

area. The area of 0.28 was at the median (approximately) of the moderate deviation area. The rest of the deviation areas were an increment of 0.03 from 0.28 through the large deviation area of 0.40.

Deviation Areas

In this study, μ and σ of the difficulty parameters (b), and μ and σ of a_1 were fixed; therefore, the variation in deviation areas was determined by the size of a_2 relative to a_1 as controlled by the weighting factor W. As W increased, a_2 and the deviation areas also increased. Because deviation areas 0.19, 0.28, 0.31, 0.34, 0.37 and 0.40 were fixed apriori, W was explored from a range of 0.34 to 0.90 to create the six deviation areas; that is, different values of W were used until a pre-specified deviation area was obtained. For each test length, all six deviation areas had the same a_1 parameters, but a_2 parameters were weighted by W.

To ensure the same deviation areas across test lengths, some minor changes were made in the a_1 parameters of the 20-item test. When the a and b parameters of the 20-item test were selected from the 40-item test (with the same mean, variance and W as the 40-item test), the average deviation areas for the 20-item test were slightly lower than the 40-item test when the deviation area was 0.40 (eg., instead of 0.40 in the 40-item test, it was 0.39). Therefore, minor changes were made in the values of a_1 (e.g, instead of the

value of a_1 for item 13 = 0.62, it was changed to 0.66). The changes were made only for the large deviation area of 0.40 (W = 0.90 as in the 40-item test) and any decrease for an item was compensated for by the same increase to another item or vice-versa (to ensure the same mean and variance). Once the changes had been made and the 20-item test had the same W, same deviation area (0.40), and about the same mean, variance, and range for a_1 as in the 40-item test, the rest of the deviation areas for the 20-item tests were weighted by the same W as the rest of the deviation areas for the 40-item tests; that is, given the same W, all the deviation areas for the 20-item tests were the same as the 40-item tests with up to 0.01 rounding errors.

To ensure the same deviation areas across the proportion of items loaded on the minor dimension, the item test parameters for the major dimension of the 10% and 20% conditions were the same as the 100% condition for each test length. For the minor dimension of the 10% and 20% condition, the same proportion of items were selected from the minor dimension of the 100% condition. Those items not loaded on the minor dimension were set to zero and only those items loaded on the minor dimension were computed for the deviation areas. Items loaded on the minor dimension of the 10% and 20% condition were selected only from the deviation area of 0.40 (of the 100% condition). After the items that averaged to about 0.40 deviation areas had been

selected (a few minor adjustments were made on the a_1 item-parameters to ensure the same deviation areas, especially for the 10% conditions), the same items were used for computing the other deviation areas, and the second dimension of the other deviation areas was weighted by the same W as used in the 100% condition. Therefore, for all deviation areas, the 10% and 20% conditions would have the same weight (W) as the 100% condition. For 10% and 20% of the items loaded on the minor dimension, the deviation areas of those items loaded were about the same as the 100% condition with rounding errors of less than 0.01.

Given a fixed deviation area, W was the same across test lengths and the proportion of items loaded on the minor dimension. In this study, the final levels of W resulting in the six deviation areas are shown in Table 3.2. For the upper limit of the minor deviation area, a_2 was about one third the size of a_1, and for the upper limit of the moderate deviation area, a_2 was about two-thirds the size of a_1.

Table 3.2

Level of W and the Six Deviation Areas

Deviation Area	0.19	0.28	0.31	0.34	0.37	0.40
W	0.34	0.54	0.62	0.70	0.80	0.90

Item Response Data Generation

The item-parameters for the two test lengths and the three proportions of items loaded on the minor dimension were used to generate item response data. The same item-parameters were used for the 700 and 1,500 sample sizes. Both the θ_1 and θ_2 were generated from a normal distribution with mean zero and variance equal to one, and θ_1 and θ_2 were independent. The means and variances of θ_1 and θ_2 were the same across replications. Two levels of sample size, two levels of test length, and three proportions of items loaded on the second dimension were crossed with each other to create 12 unique conditions. Table 3.3 presents the design for sample size, test length, and proportion of items loaded on the minor dimension. The generation of item responses was repeated 100 times for each of the six deviation areas, totaling 7200 data sets. Table 3.4 shows replication of the three categories of error based on fixed deviation areas for the 12 conditions.

Test of Hypothesis

For each item response data set generated, Stout's nonparametric procedure was used to test the hypothesis of H_o: $d = 1$ versus H_1: $d > 1$; that is, whether the data were essentially unidimensional. Stout's (1991) dimensionality testing program (DIMTEST) was used. In DIMTEST, AT1 items can be selected either by expert's opinion or by factor analysis of tetrachoric correlations (see Chapter 2). In

Table 3.3

Design for Sample Size, Test Length, and Proportion of Items Loaded on the Second Dimension

Sample Size	Test Length	Prop. of Items	Condition
700	20	10%	1
		20%	2
		100%	3
	40	10%	4
		20%	5
		100%	6
1,500	20	10%	7
		20%	8
		100%	9
	40	10%	10
		20%	11
		100%	12

Table 3.4

Replication of the Three Categories of Error for the 12 Conditions

		Conditions		
Category	Area	1	2	3.....12
Minor	0.19	100	100	100
Moderate	0.28	100	100	100
	0.31	100	100	100
	0.34	100	100	100
Large	0.37	100	100	100
	0.40	100	100	100

this study, factor analysis was used and the sample size used in factor analysis was 700 and 1,500 (same sample size as DIMTEST). Because the purpose of factor analysis was to select items for AT1 in DIMTEST, 10 factor analyses were performed for each unique condition (using different replications) and the factor analysis that produced the most dimensionally distinct items (as determined from examining the item parameters) for AT1 was used for the rest of the replications; that is, the same AT1 items were used for 100 replications. Thus, differences across the replications could not be attributed to the use of different item parameters (selected for AT1) being used in Stout's test of essential unidimensionality.

Simulation Models

Both the univariate 2PL model and the bivariate extension of the 2PL model with compensatory abilities were used in the generation of data. Two dimensional items were generated using the following equation:

$$P_i(\theta_1,\theta_2) = \frac{1}{1+\exp[-1.7[a_{1i}(\theta_1-b_{1i}) + a_{2i}(\theta_2 - b_{2i})]]} \quad (3.5)$$

where

θ_1 and θ_2 are the ability parameters for dimensions one and two,

a_{1i} and a_{2i} are the discrimination parameters for item i on the two dimensions, and

b_{1i} and b_{2i} are the difficulty parameters for item i.

Nandakumar (1991) showed that when a_2 and b_2 of the second dimension are zero, equation 3.5 reduces to a unidimensional 2PL logistic model with respect to θ_1. Therefore, when 10% and 20% of the items were two-dimensional, unidimensional items were simulated by using the unidimensional 2PL model

$$P_i(\theta_1) = \frac{1}{1+\exp(-1.7a_i(\theta_1-b_i))} \tag{3.6}$$

Analysis of Data

The analysis of variance was used to investigate the effect of deviation area, test length, sample size, and the proportion of items loaded on the minor dimension on the rejection rates of Stout's procedure for every category of deviations. The significance level for each of the analyses was set at .05. The model for the analysis of variance was

$$\begin{aligned} Y_{ijkl} = \mu + \kappa_h + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \\ \kappa\alpha_{hi} + \kappa\beta_{hj} + \kappa\gamma_{hk} + \kappa\alpha\beta_{hij} + \kappa\alpha\gamma_{hik} + \\ \alpha\beta\gamma_{ijk} + \kappa\beta\gamma_{hjk} + \epsilon_{hijkl} \end{aligned} \tag{3.7}$$

where

Y_{ijkl} = rejection rates based on Stout's procedure
μ = grand mean
κ_h = effect of deviation areas, h = 1, 2, 3
α_i = effect of sample size, i = 1, 2
β_j = effect of test length, j = 1, 2

The rejection rate for each condition was also compared with each unique area to assess if Stout's procedure was rejected at a rate directly proportional to the estimation errors. The power curves would be used to describe the

power of Stout's procedure under variations of estimation errors.

Summary

The purpose of this study was to determine the power of Stout's procedure in assessing essential unidimensionality under minor, moderate, and large deviations from unidimensionality. The purpose and research questions, the design of the study, the simulation model, and the analysis of data were presented.

CHAPTER 4
RESULTS

The purpose of the study was to investigate the power of Stout's procedure to detect deviations from essential unidimensionality in two-dimensional data. The conditions manipulated were the deviation area, test length, sample size, and the proportion of items loaded on the minor dimension. Table 4.1 shows the distribution of rejection rates across all conditions.

The analysis of variance was used to investigate the effect of deviation area, test length, sample size, and the proportion of items loaded on the minor dimension on the power of Stout's procedure. The significance level for each of the analyses was set at .05. Table 4.2 summarizes the results of analysis of variance of the rejection rates of Stout's test of essential unidimensionality (the rejection rate is the dependent variable) across four factors: deviation area (A), proportion of items loaded on the minor dimension (P), test length (L), and sample size (S). None of the three-way interactions listed in Table 4.2 were significant. Among the six possible two-way interactions, only AxL, PxS, and PxL were significant. The variance explained by the deviation area by test length (θ_{al}^2=36.02)

Table 4.1

The Distribution of Rejection Rates Across All Conditions.

		Sample Size			
		700		1,500	
Areas	Proportion	20*	40*	20*	40*
0.19	100%	19	11	34	36
	20%	19	30	25	52
	10%	4	13	24	32
0.28	100%	26	28	43	68
	20%	64	90	91	98
	10%	19	57	29	88
0.31	100%	24	41	43	82
	20%	66	92	99	98
	10%	15	75	29	94
0.34	100%	35	59	58	99
	20%	87	99	99	100
	10%	19	87	44	98
0.37	100%	34	80	63	100
	20%	97	100	100	100
	10%	27	99	60	100
0.40	100%	53	88	76	100
	20%	100	100	100	100
	10%	27	100	62	100

Note. * refers to test length.

Table 4.2

Analysis of Variance for Sample Size, Test Length, Degree of Deviation from Unidimensionality, Proportion of Items Loaded on the Minor Dimension, and Their Interactions.

Source of Variation	df	Sum of Squares	F	P	Variance Explained
Deviation Area(A)	5	27626.16	72.81	0.0001	454.11
Proportion(P)	2	13806.33	90.97	0.0001	284.47
Size of Examinees(S)	1	5688.88	74.97	0.0001	155.91
Test Length(L)	1	13338.88	175.78	0.0001	368.41
AP	10	2079.50	2.74	0.0637	33.01
AS	5	261.27	0.69	0.6433	0.00
AL	5	1460.27	3.85	0.0332	36.02
PS	2	722.11	4.76	0.0353	23.76
PL	2	4755.11	31.33	0.0001	191.80
LS	1	32.00	0.42	0.5307	0.00
APS	10	257.72	0.34	0.9483	0.00
APL	10	2171.72	2.86	0.0562	70.64
ALS	5	415.16	1.09	0.4206	2.38
PLS	2	292.00	1.92	0.1964	11.68
Error	10	758.83			75.88

Note. The dependent variable is the rejection rates per 100 trials based on Stout's hypothesis test of essential unidimensionality.

interaction and the proportion of items by sample size (θ_{ps}^2=23.76) interaction were relatively small compared to other factors; they were 2% and 1.4% of the total variance, respectively. The variance for the proportion of items by test length interaction (θ_{pl}^2=191.80), however, was moderately high, which explained about 11% of the total variance in the rejection rates. All the main effects (test length, proportion of items, sample size and deviation area) were significant.

Interaction Effects

PxL Interaction

The interaction between the proportion of items loaded on the minor dimension and test length is shown in Table 4.3. The difference in power (of Stout's procedure) between

Table 4.3

Mean for the Interaction of Proportion of Items (P) and Test Length

	Test Length	
P	20	40
100%	42.33	66.00
20%	78.91	88.25
10%	29.92	78.58

20-item tests and 40-item tests was greater at the 10% condition than at the 20% and 100% conditions, but the difference in power between 20-item tests and 40-item tests was greater at the 100% condition than at the 20% condition. The differences in power among the three proportions were consistently lower for a 40-item test than for a 20-item test.

AxL Interaction

The interaction between deviation area and test length is shown in Table 4.4. The difference in power (of Stout's procedure) between 20-item tests and 40-item tests was lower at a minor deviation area (0.19) than at moderate or large

Table 4.4

Mean for the Interaction of Deviation Area (A) and Test Length

	Test Length	
A	20	40
0.19	20.83	29.00
0.28	45.33	71.50
0.31	46.00	80.33
0.34	57.00	90.33
0.37	63.50	96.50
0.40	69.66	98.00

deviation areas. Also, the increase in power across deviation areas was much greater for 40-item tests than for 20-item tests.

PxS Interaction

The interaction between the proportion of items loaded on the minor dimension and sample size is shown in Table 4.5. The difference in power (of Stout's procedure) between the 1,500 and 700 sample size conditions was higher at the 100% condition than at the 10% and 20% conditions, and the difference in power between the 700 and 1,500 sample size conditions was higher at the 10% condition than at the 20% condition. Also, the difference in power across the three proportions was lower for the 1,500 sample size than for the 700 sample size.

Table 4.5

Mean for the Interaction of Proportion of Items (P) and Sample Size

	Sample Size	
P	700	1,500
100%	41.50	66.83
20%	78.66	88.50
10%	45.16	63.33

Main Effects

Deviation Areas

As shown in Table 4.2, there was a significant main effect due to deviation area. Using Duncan's multiple range test, Table 4.6 shows that on the average, differences in rejection rates existed between the 0.19 deviation area (minor deviation) and the rest of the deviation areas (moderate and large). Differences in rejection rates also exist between 0.28, 0.31 and 0.34, 0.37, 0.40, and between 0.34 and 0.37, 0.40.

As shown in Table 4.6, the average rejection rate for the minor deviation was about 25%. Table 4.1 shows that regardless of sample size, proportion of items loaded on the second dimension, and test length, none of the rejection rates for minor deviations were above 52% and the lowest rejection rate was 4%.

Table 4.6

Mean Rejection Rates for Each Deviation Area Based on Duncan's Multiple Range Test

Deviation Area					
0.40	0.37	0.34	0.31	0.28	0.19
83.83^{a}	80.00^{a}	73.67^{b}	63.17^{c}	58.42^{c}	24.92^{d}

Note. The means with different superscripts differ significantly at $p < .05$.

For the moderate deviations, the rejection rates averaged about 65% (Table 4.6). Table 4.1 shows that the rejection rates for the moderate deviations ranged from 19% to 100%. For the large deviations, the average rejection rate was 82%, and the rejection rates ranged from 27% to 100%.

Effect of Test Length

As shown in Table 4.2, there was a significant difference between rejections for the two test lengths. As shown in Table 4.7, 20-item tests averaged a 50% rejection rate and 40-item tests averaged a 78% rejection rate.

Table 4.7

Mean Rejection Rates for Each Test Length

Test Length	
20	40
50.39	77.61

Proportion of Items

As shown in Table 4.2, there was a significant difference between the rejection rates obtained with different proportions of items loaded on the minor dimension. Pairwise comparisons using Duncan's multiple range test in Table 4.8 show that differences in rejection rates occured between the 20% and 100% conditions, and

between the 20% and 10% conditions. Also, the 10% and 100% conditions averaged about 54% rejection rates, but the 20% condition averaged about 84%.

Table 4.8

Mean Rejection Rates for Each Proportion of Items Based on Duncan's Multiple Range Test

Proportion of Items		
100%	20%	10%
54.250^{a}	83.583^{b}	54.167^{a}

Note. The mean with different superscripts differ significantly at $p < .05$.

Sample Size

As shown in Table 4.2, there is a significant difference in rejection rates between the two sample sizes. Table 4.9 shows that the 700 subject condition averaged

Table 4.9

Mean Rejection Rates for Each Sample Size

Sample Size	
700	1,500
55.11	72.89

about a 55% rejection rate and the 1,500 subject condition averaged about a 73% rejection rate.

The Power of Stout's Procedure

The power curves from Figures 4.1 through 4.12 show that for all test lengths, sample sizes, and proportions of items loaded on the minor dimension, the power increased as the deviation area increased. In general, the rejection rates for each condition were directly related to the size of the deviation areas.

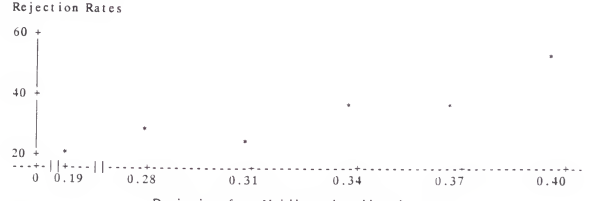

Figure 4.1

Power Curves for Deviations from Unidimensionality: Sample Size = 700, Proportion (on 2nd Dimension) = 100% and Test Length = 20

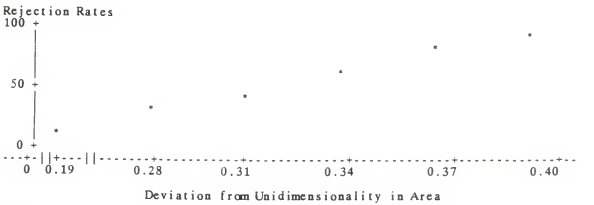

Figure 4.2

Power Curves for Deviations from Unidimensionality: Sample Size = 700, Proportion = 100% and Test Length = 40

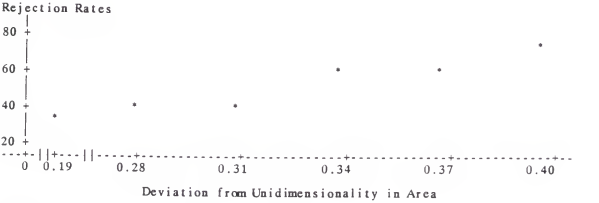

Figure 4.3

Power Curves for Deviations from Unidimensionality: Sample Size = 1,500, Proportion = 100% and Test Length = 20

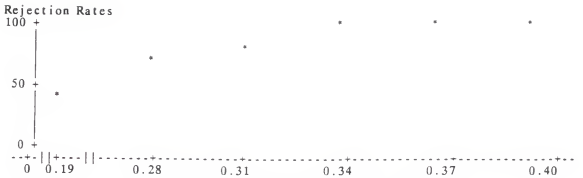

Figure 4.4

Power Curves for Deviations from Unidimensionality: Sample Size = 1,500, Proportion = 100% and Test Length = 40

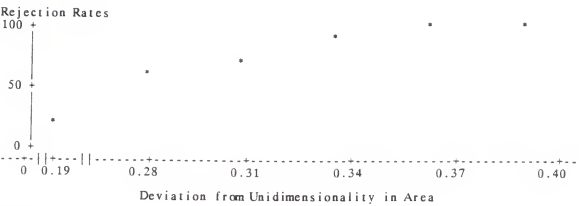

Figure 4.5

Power Curves for Deviations from Unidimensionality: Sample Size = 700, Proportion = 20% and Test Length = 20

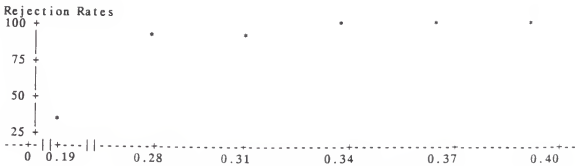

Figure 4.6

Power Curves for Deviations from Unidimensionality: Sample Size = 700, Proportion = 20% and Test Length = 40

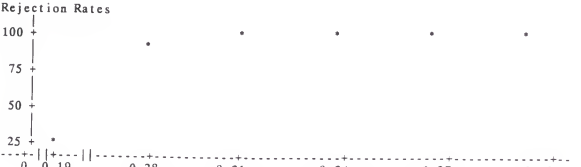

Figure 4.7

Power Curves for Deviations from Unidimensionality: Sample Size = 1,500, Proportion = 20% and Test Length = 20

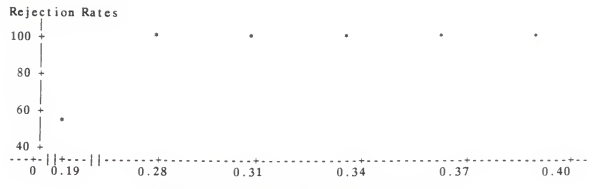

Figure 4.8

Power Curves for Deviations from Unidimensionality: Sample Size = 1,500, Proportion = 20% and Test Length = 40

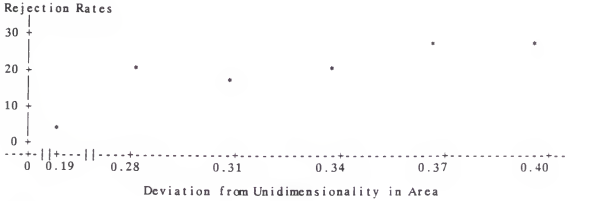

Figure 4.9

Power Curves for Deviations from Unidimensionality: Sample Size = 700, Proportion = 10% and Test Length = 20

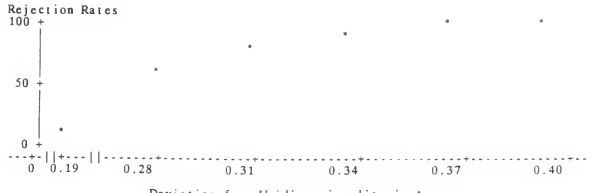

Figure 4.10

Power Curves for Deviations from Unidimensionality: Sample Size=700, Proportion = 10% and Test Length = 40

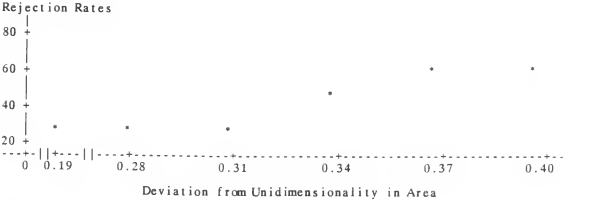

Figure 4.11

Power Curves for Deviations from Unidimensionality: Sample Size=1,500, Proportion = 10% and Test Length = 20

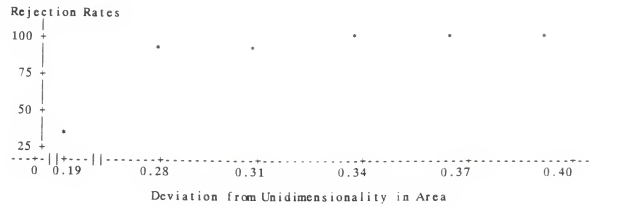

Figure 4.12

Power Curves for Deviations from Unidimensionality: Sample Size=1,500, Proportion = 10% and Test Length = 40

CHAPTER 5
DISCUSSION AND CONCLUSION

Discussion

The results of this study are discussed in relation to the correlation between the ability of the major dimension (θ_1) and the reference composite ability (γ) investigated in the preliminary investigation. In this investigation, the population correlation between γ and θ_1 was computed for each combination of deviation area, test length, and proportion of items on the second dimension. The population correlation between the reference composite (γ) and θ_1

$$\rho_{\gamma,\theta_1} = \frac{W_1}{(W_1^2 + W_2^2)^{0.5}} \tag{5.1}$$

where
w_1(p x 1) is the first eigenvector of the matrix A'A, and
w_2(p x 1) is the second eigenvector of the matrix A'A

was derived from the estimated reference composite of Wang (1986),

$$\hat{\gamma}_j = \theta_j W_1 \tag{5.2}$$

where
θ_j(1 x p) is the jth row of the matrix θ, and
w_1(p x 1) is the first eigenvector of the matrix A'A.

A low correlation means that the data were heavily influenced by the minor ability and the ability of interest

(the major dimension) did not correspond to the reference composite. A high correlation means the influence of the minor ability was very mild and the ability of interest (major dimension) was consistent with the reference composite.

Proportion of Items

When the proportion of items loaded on the second dimension was 100%, the results of a preliminary investigation in Table 5.0 showed that the correlations between θ_1 of the simulated data and the reference composite were inversely related to the size of the deviation areas; that is, with both test lengths, high deviation areas yielded lower correlations and low deviation areas yielded higher correlations. The rejection rates were also directly related to the deviation areas. Looking at each of the deviation areas for 100% of the items loaded on the second dimension, when the correlation was high, such as 0.954, Stout's procedure rejected on the average of 27% for 20-item tests and 24% for 40-item tests. When the correlation was 0.75, the rejection rate of Stout's procedure was 65% on the average for 20-item tests and 94% for 40-item tests.

For the 20% of the items loaded on the second dimension condition, the results in Table 5.0 show that the correlations between theta 1 of the simulated data and the reference composite remained near the 0.99 or 0.98 levels, regardless of the deviation area or test length. Although

Table 5.0

Correlation Between Theta 1 and the Reference Composite

Test Length	% of Items	Deviation Areas 0.19	0.28	0.31	0.34	0.37	0.40
20	100	0.954 (27%)	0.891 (35%)	0.861 (34%)	0.829 (47%)	0.788 (49%)	0.747 (65%)
	20	0.998 (22%)	0.994 (78%)	0.992 (83%)	0.990 (93%)	0.986 (99%)	0.982 (100%)
	10	0.999 (14%)	0.999 (24%)	0.998 (22%)	0.997 (32%)	0.996 (44%)	0.995 (45%)
40	100	0.954 (24%)	0.891 (48%)	0.861 (62%)	0.829 (79%)	0.788 (90%)	0.747 (94%)
	20	0.997 (41%)	0.994 (94%)	0.991 (94%)	0.989 (99%)	0.985 (100%)	0.980 (100%)
	10	0.999 (23%)	0.998 (73%)	0.997 (85%)	0.996 (93%)	0.995 (100%)	0.994 (100%)

Note. The value in each of the parentheses is the average rates of the two sample sizes corresponding to the correlation $\rho_{Y,\theta 1}$.

the rejection rates were related to the deviation areas, the relationship between the correlation $\rho_{Y,\theta 1}$ and the deviation area was very mild. When $\rho_{Y,\theta 1}$ was very high over all the deviation areas, Stout's null hypothesis of essential unidimensionality was less powerful. In this study, however, when 20% of the items loaded on the minor trait, the rejection rates for minor, moderate, and large deviation areas were very high and some of the rejection rates were

100%. One reason for these high rejection rates was the selection of items into subtest AT1 through factor analysis. Factor analysis selects the M items into AT1 that load most heavily either positively or negatively on the second extracted factor (i.e., the selected items are dimensionally distinct from the rest of the items), resulting in the possible selection of most of the 20% ($M \leq N/4$) items that loaded on the second factor. Because the rest of the items were not loaded on the second factor, the selection of items for AT2 might not have had the same difficulty distribution as in AT1. Thus, examinees within each subgroup of PT were not likely to be approximately equal on the dominant trait measured by the test, which resulted in high rejection rates.

Similar to the results obtained with the 20% condition, the correlations between theta 1 of the simulated data and the reference composite of the 10% condition were at the 0.99 level regardless of the deviation area or test length. The rejection rates for the 10% condition, however, were also relatively high. The high rejection rates for the 10% condition might be the result of the same factor as the high rejection rates for the 20% condition.

Although the rejection rates for the 20% condition were higher than for the conditions in which 100% of the items loaded on the second dimension, the correlation $\rho_{Y,\theta 1}$ for the 20% condition was much higher than for the 100% condition.

This shows that the high rejection rates for the 20% condition (and for the 10% condition) might be the result of the weakness of Stout's procedure in selecting dimensionally distinct AT1 items in DIMTEST. Because the scores on AT1 are used to compute Stout's statistics (see Chapter 2), if AT1 items are dimensionally distinct from the rest of the items, the null hypothesis of Stout's procedure will be rejected. This can be a problem. If no dimensionally distinct items are present when all items load on both dimensions, Stout's null hypothesis may not be rejected. If there is a small proportion of dimensionally distinct items such as 10%, even when the weight on the second dimension is very weak, the null hypothesis may be rejected.

Test Length

Although the preliminary investigation showed that the two test lengths have about the same levels of correlation $\rho_{\gamma,\theta 1}$ over deviation areas, the finding of the present study suggests that Stout's procedure has more power in rejecting the null hypothesis of essential unidimensionality with longer tests than with shorter tests; that is, the power increased from the 20-item tests to the 40-item tests under moderate and large deviation areas, sample sizes, and proportions of items loaded on the second dimension. Although for minor deviation areas with 100% of the items loaded on the minor dimension, the 20-item test had slightly more power than the 40-item test, the result might be due to

random error (the increase was very mild). In general, the results of this study are consistent with Nandakumar's observation (1991).

Sample Size

As shown in this study, when the sample size increased, the power of Stout's procedure also increased. The results were consistent with those of Stout (1987) and Nandakumar (1991). Larger sample sizes not only had an advantage over smaller sample sizes in terms of power using Stout's procedure, but larger sample sizes also provided a more stable estimation in factor analysis (Gorsuch, 1983) and thus, factor analysis selected better dimensionally distinct items for AT1 in DIMTEST.

Deviation Areas

In general, for all proportions of items loaded on the second dimension, test lengths, and sample sizes, as the deviation area increased, the rejection rate also increased; that is, the rejection rate for each condition is directly related to the deviation area. Although the rejection rates were directly related to deviation areas, μ, σ^2 and the range of a_s and b_s were fixed in this study. The impact of μ, σ^2 and the range of a_s and b_s on deviation areas and the power of Stout's procedure need to be further explored with variations in these parameters.

The Power of Stout's Procedure

Although factor analysis is merely a data analytic technique for obtaining AT1 items that are as dimensionally distinct from the rest of the items as possible (Stout, 1987), a preliminary study found that when the deviation areas were minor and moderate, the power of Stout's procedure fluctuated as a function of the items selected by the factor analysis for AT1; that is, the more dimensionally distinct AT1 items tended to have higher rejection rates. Also, factor analysis did not lead to selection of the most dimensionally distinct items for AT1 when the sample size used was small (500) and the test length was 40 items. To avoid the possibility that AT1 items selected by factor analysis might not be the most dimensionally distinct items, factor analysis was performed on the sample sizes of 700 and 1,500 (same as in DIMTEST) and attempts were made to ensure that items selected by factor analysis for AT1 were as dimensionally distinct from the rest of the items as possible; that is, 10 factor analyses were performed on the data sets for each condition and the factor analysis that yielded the most dimensionally distinct items for AT1 was used. In this study, given a fixed condition, the same AT1 items (the most dimensionally distinct set of items yielded by factor analysis) were used for 100 replications.

The results of this study showed that the power of Stout's procedure in rejecting the null hypothesis of

essential unidimensionality was conditioned on sample size, test length, proportion of items loaded on the second dimension, and deviation area. The power for each condition was directly related to the deviation area; that is, the larger the deviation area, the greater the power. A sample size of 1,500 had more power than a sample size of 700, and a 40-item test had more power than a 20-item test. In general, the power of Stout's procedure was relatively low for 20-item tests with 700 examinees but relatively high for 40-item tests with 1,500 examinees and this was true for all proportions of items loaded on the minor dimension.

Data Appropriate for Unidimensional IRT Estimation

Stout (1987) and Nandakumar (1991) have assumed that essentially unidimensional data are appropriate for unidimensional IRT estimation. In this study, data in the minor deviation area conditions were supposed to be appropriate for unidimensional IRT estimation. The rejection rates for minor deviations were higher than the 5% nominal level. Therefore, either Stout's procedure had too much power in rejecting the null hypothesis of essential unidimensionality or the cut-off for minor deviations needs to be reexamined.

Comparison to Previous Studies

Although Stout (1987) studied strictly unidimensional data and two dimensional data of equal weight (two equal dominant dimensions), this study only examined two-

dimensional data with one major and one minor dimension. When the weight of the second dimension was large (large deviation areas), the power in this study was comparable to the results of Stout's (1987) two-dimensional data.

Nandakumar (1991) also examined the power of Stout's test of essential unidimensionality. But, the weight of the second dimension in this study was based on the weighting factor W, as opposed to ξ in Nandakumar (see Chapter 3); that is, the distributions of a_1 and b_1 of the major dimension were the same regardless of the weight of the second dimension. Because the major dimension was kept constant, there was no confounding of σ_{a1} and μ_{a1} across deviation areas. In contrast, Nandakumar (1991) generated data where σ_{a1} and μ_{a1} decreased as ξ increased. Because there was no reduction of σ_{a1} and μ_{a1} across deviation areas, the power in this study was much higher than the power in Nandakumar's study. The item parameters in this study were fixed across conditions, thus the results are easier to interpret across sample size and the proportion of items loaded on the second dimension than in Nanadakumar (1991).

Limitations of the Present Study

The criteria used in determining the three levels of deviations from unidimensionality were based on the criteria used by Shepard et al. (1985) in the categorization of biases between two groups. The results and conclusions of

this study were, therefore, limited to the accuracy of these criteria.

Although two-dimensional data were used in the present study, the dimensions were assumed to be uncorrelated and guessing was not taken into account. Other correlations between dimensions, other μ, σ^2 and ranges of a_s and b_s, other sample sizes, other test lengths, and other proportions of items loaded on the second dimension might lead to different results.

In this study, many factor analyses were performed for each condition to ensure that items selected for AT1 were as dimensionally distinct from the rest of the items as possible. Since known items parameters were also used in the selection of the AT1 items, the results may have more power than in an applied data. In practice with real data, this may not be possible because the data set may not be large enough to perform many factor analyses. Therefore, when working with real data, experts' opinions may be used in selecting data when appropriate. If factor analysis is used, care should be undertaken to ensure that the AT1 items are dimensionally distinct from the rest of the items by ensuring that only the highest a_1 with the lowest a_2, or vice versa, are chosen for AT1 items (Stout, personal communication).

Conclusions

The results of this study indicated that the power of Stout's procedure is directly related to the deviation areas. Deviation areas were inversely proportional to the correlation between the dominant ability and the reference composite. When the sample size increased, the power of Stout's procedure also increased. The power of Stout's procedure for 40-item tests was higher than for 20-item tests. When the proportion of items loaded on the minor dimension was 20%, the power was the highest. Although the power for the 20% condition was higher than for the 100% condition, the correlation $\rho_{\gamma,\theta 1}$ for the 20% condition was also extremely high. For the 20% and 10% conditions, even when the correlations $\rho_{\gamma,\theta 1}$ is near 1.00, the rejection rates can be high.

In general, Stout's (1987) procedure had sufficient power in rejecting the null hypothesis of essential unidimensionality if the combination of σ_{a1}^2 and μ_{a1}, and σ_{a2}^2 and μ_{a2} were such that about 10% to 20% of the items selected into AT1 (under all conditions) were dimensionally distinct from the rest of the items. If AT1 was dimensionally distinct from the rest of the items, then Stout's null hypothesis of essential unidimensionality would be rejected.

The results of this study indicate that for minor deviation areas, the rejection rates of Stout's procedure

were not near the nominal level of 5%. For moderate and large deviation areas, Stout's (1987) procedure had sufficient power in rejecting the null hypothesis of essential unidimensionality, especially if the sample size was 1,500 and the test length was 40. The appropriateness of essential unidimensional data for unidimensional IRT estimation is unknown.

Further Research

Although there is a direct relationship between deviation areas and rejection rates, the relationship between deviation areas and $\rho_{\gamma,\theta 1}$ implies that the cut-off points for minor, moderate and large deviation areas may not be appropriate. Therefore, more studies need to be undertaken to check if the cut-off points based on Shepard et al. (1985) and used in this study for the three deviation areas were indeed appropriate cut-off points for unidimensional IRT estimation. The impact of μ, σ^2 and the range of a_s and b_s on deviation areas also need to be further explored.

The results of this study and $\rho_{\gamma,\theta 1}$ imply that Stout's procedure may be too powerful; therefore, some adjustments in Stout's procedure need to be undertaken. A test for essentially unidimensional data could become a test for the appropriateness of unidimensional IRT estimation with two-dimensional data.

The appropriateness of essentially unidimensional data for equating and adaptive testing has not been explored. Other variables that may influence the power of Stout's procedure, such as the direction of items and guessing, may need to be systematically studied. Finally, only one major and one minor ability were studied here. Preliminary investigation showed Stout's procedure had more power when more than one minor ability was present. Therefore, the power of Stout's procedure based on one major and many minor abilities may need further research.

REFERENCES

Ackerman, T. A. (1987, April). The use of unidimensional item parameter estimations of multidimensional items in adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. Applied Psychological Measurement, 13, 113-127.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-296.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R.K. Hambleton (Ed.), Application of item response theory (pp. 1-20). Vancouver, BC: Educational Research Institute of British Columbia.

Ben-Simon, A., & Cohen, Y. (1990, April). Rosenbaum's test of unidimensionality: Sensitivity analysis. Paper presented at the annual meeting of the American Education Research Association, Boston.

Berger, M., & Knol, D. (1990, April). On assessment of dimensionality in multidimensional item response theory models. Paper presented at the annual meeting of the American Education Research Association, Boston.

Bock, R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-458.

Bock, R., Gibbons, R., & Muraki, E.(1988). Full-information item factor analysis. Applied Psychological Measurement, 12, 261-280.

Christofferson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Drasgow, F. (1987). A study of measurement bias of two standard psychological tests. Journal of Applied Psychology, 72, 19-30.

Drasgow, F., & Parsons, C. (1983, April). The analysis of dichotomous test data using factor analytic methodology. Paper presented at the annual meeting of the American Education Research Association, Montreal.

Folk, V., & Green, B. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. Applied Psychological Measurement, 13, 373-389.

Gorsuch, R. (1983). Factor analysis. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Hambleton, R. K., & Rovinelli, R. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.

Hambleton R. K., & Swaminathan, H. (1985). Item response theory: Principle and application. Boston: Kluwer Nijhoff.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violation of the unidimensionality assumption. Journal of Educational Statistics, 11, 91-115.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. Annals of Statistics, 14, 1523-1543.

Hulin, C., Drasgow, F., & Parsons, C. (1983). Item response theory: Application to psychological measurement. Homeland, IL: Dow Jones-Irwin.

Humphreys, L. (1986). An analysis and evaluation of test and item bias in the prediction context. Journal of Applied Psychology, 71, 327-333.

Kingsbury, G. (1985, April). A comparison of item response theory procedures for assessing response dimensionality. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Kingston, N., & McKinley, R. (1988, April). Assessing the structure of the GRE general test using confirmatory multidimensional item response theory. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Lawrence, I., & Dorans, N. (1987, April). An assessment of the dimensionality of SAT-Mathematics. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Lord, F. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F., & Novick, M. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

McDonald, R. (1967). Nonlinear factor analysis. British Journal of Mathematical and Statistical Psychology, 20, 205-215.

McDonald, R. (1981). Dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

McDonald, R. (1982). Linear versus non-linear models in item response theory. Applied Psychological Measurement, 6, 379-396.

McDonald, R. (1985). Factor analysis and related methods. Hillsdale, NJ: Lawrence Erlbaum.

McKinley, R. (1983, April). A multidimensional extension of the two parameter logistic latent trait model. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Mislevy, R. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Muraki, E., & Engelhard, G. (1985). Full information item factor analysis: Application of EAP scores. Applied Psychological Measurement, 9, 417-430.

Muthen, B. (1978). Contribution to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

Nandakumar, R. (1987). Refinement of Stout's procedure for assessing latent trait unidimensionality (Doctoral dissertation, University of Illinois at Urbana-Champaign, 1987). Dissertation Abstracts International, 49, 01A.

Nandakumar, R. (1991). Traditional dimensionality vs. essential dimensionality. Journal of Educational Measurement, 28, 1-19.

Nandakumar, R., & Stout, W. F. (in press). Refinement of Stout's procedure for assessing latent trait unidimensionality. Journal of Educational Statistics.

Raju, N. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

Reckase, M. (1981, April). Guessing and dimensionality: The search for a unidimensional latent space. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Reckase, M. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

Reckase, M. (1990, April). Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Reckase, M., & Ackerman, T. (1986, April). Building a test using items that require more than one skill to determine a correct answer. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Reckase, M., Ackerman T., & Carlson J.E. (1988). Building a unidimensional test using multidimensional items. Journal of Educational Measurement, 25, 193-203.

Reckase, M., & McKinley, R. (1983). Some latent trait theory in a multidimensional space. In D.J. Weiss (Ed.), Proceedings of the 1982 item response theory/computer adaptive testing conference (pp. 157-177). Minneapolis: University of Minnesota, Department of Psychology.

Salih, F. (1987). An empirical evaluation of full-information item factor analysis (Doctoral dissertation, University of Iowa, 1987). Dissertation Abstracts International, 48, 10A.

Shepard, L., Camilli, G., & Williams, D. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. (1990). A new item response theory modeling approach with application to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Stout, W., Nandakumar, R., Junker, B., Chang, H-H., & Steidinger, D. (1991). DIMTEST and TESTSIM [computer application software]. Urbana-Champaign: University of Illinois, Dept. of Statistics.

Traub, R. (1983). A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.), Application of item response theory (pp. 57-70). Vancouver: Educational Research Institute of British Columbia.

Wang, M. (1986, April). Fitting a unidimensional model to multidimensional item response data. Paper presented at the ONR contractors conference, Gatlinburg, TN.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. Journal of Educational Measurement, 24, 293-308.

BIOGRAPHICAL SKETCH

Cheng Ang was born on December 28, 1960, in Malacca, Malaysia. He graduated with a Bachelor of Arts degree from Loma-Linda University in 1983 and a Master of Arts degree from Andrews University in 1984. In 1986 he was certified as a high school mathematics teacher in Michigan. In the spring of 1990 he started the Ph.D. program in educational research and evaluation methodology at the University of Florida.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Michael Miller, Chair
Associate Professor of
Foundations of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James Algina
Professor of Foundations of
Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Linda Crocker
Professor of Foundations of
Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Clemens Hallman
Professor of
Instruction and Curriculum

This dissertation was submitted to the Graduate Faculty of the College of Education and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 1992

Chairman, Foundations of Education

Dean, College of Education

Dean, Graduate School